Imputation (statistics)

In statistics, **imputation** is the process of replacing missing data with substituted values. When substituting for a data point, it is known as "**unit imputation**"; when substituting for a component of a data point, it is known as "**item imputation**". There are three main problems that missing data causes: missing data can introduce a substantial amount of bias, make the handling and analysis of the data more arduous, and create reductions in efficiency.[1] Because missing data can create problems for analyzing data, imputation is seen as a way to avoid pitfalls involved with listwise deletion of cases that have missing values. That is to say, when one or more values are missing for a case, most statistical packages default to discarding any case that has a missing value, which may introduce bias or affect the representativeness of the results. Imputation preserves all cases by replacing missing data with an estimated value based on other available information. Once all missing values have been imputed, the data set can then be analysed using standard techniques for complete data.[2] There have been many theories embraced by scientists to account for missing data but the majority of them introduce bias. A few of the well known attempts to deal with missing data include: hot deck and cold deck imputation; listwise and pairwise deletion; mean imputation; non-negative matrix factorization; regression imputation; last observation carried forward; stochastic imputation; and multiple imputation.

Contents

Listwise (complete case) deletion

By far, the most common means of dealing with missing data is listwise deletion (also known as complete case), which is when all cases with a missing value are deleted. If the data are missing completely at random, then listwise deletion does not add any bias, but it does decrease the power of the analysis by decreasing the effective sample size. For example, if 1000 cases are collected but 80 have missing values, the effective sample size after listwise deletion is 920. If the cases are not missing completely at random, then listwise deletion will introduce bias because the sub-sample of cases represented by the missing data are not representative of the original sample (and if the original sample was itself a representative sample of a population, the complete cases are not representative of that population either).[3] While listwise deletion is unbiased when the missing data is missing completely at random, this is rarely the case in actuality.[4]

Pairwise deletion (or "available case analysis") involves deleting a case when it is missing a variable required for a particular analysis, but including that case in analyses for which all required variables are present. When pairwise deletion is used, the total N for analysis will not be consistent across parameter estimations. Because of the incomplete N values at some points in time, while still maintaining complete case comparison for other parameters, pairwise deletion can introduce impossible mathematical situations such as correlations that are over 100%.[5]

The one advantage complete case deletion has over other methods is that it is straightforward and easy to implement. This is a large reason why complete case is the most popular method of handling missing data in spite of the many disadvantages it has.

Single imputation

Hot-deck

A once-common method of imputation was hot-deck imputation where a missing value was imputed from a randomly selected similar record. The term "hot deck" dates back to the storage of data on punched cards, and indicates that the information donors come from the same dataset as the recipients. The stack of cards was "hot" because it was currently being processed.

One form of hot-deck imputation is called "last observation carried forward" (or LOCF for short), which involves sorting a dataset according to any of a number of variables, thus creating an ordered dataset. The technique then finds the first missing value and uses the cell value immediately prior to the data that are missing to impute the missing value. The process is repeated for the next cell with a missing value until all missing values have been imputed. In the common scenario in which the cases are repeated measurements of a variable for a person or other entity, this represents the belief that if a measurement is missing, the best guess is that it hasn't changed from the last time it was measured. This method is known to increase risk of increasing bias and potentially false conclusions. For this reason LOCF is not recommended for use.[6]

Cold-deck

Cold-deck imputation, by contrast, selects donors from another dataset. Due to advances in computer power, more sophisticated methods of imputation have generally superseded the original random and sorted hot deck imputation techniques. It is a method of replacing with response values of similar items in past surveys. It is available in surveys that measure time intervals.

Mean substitution

Another imputation technique involves replacing any missing value with the mean of that variable for all other cases, which has the benefit of not changing the sample mean for that variable. However, mean imputation attenuates any correlations involving the variable(s) that are imputed. This is because, in cases with imputation, there is guaranteed to be no relationship between the imputed variable and any other measured variables. Thus, mean imputation has some attractive properties for univariate analysis but becomes problematic for multivariate analysis.

Mean imputation can be carried out within classes (i.e. categories such as gender), and can be expressed as $\hat{y}_i = \bar{y}_h$ where $\hat{y}_i$ is the imputed value for record i and $\bar{y}_h$ is the sample mean of respondent data within some class h. This is a special case of generalized regression imputation:

$$\hat{y}_{mi} = b_{r0} + \sum_j b_{rj} z_{mij} + \hat{e}_{mi}$$

Here the values b_{r0}, b_{rj} are estimated from regressing y on x in non-imputed data, z is a dummy variable for class membership, and data are split into respondent (r) and missing (m).[7][8]

Non-negative matrix factorization

Non-negative matrix factorization (NMF) can take missing data while minimizing its cost function, rather than treating these missing data as zeros that could introduce biases.[9] This makes it a mathematically proven method for data imputation. NMF can ignore missing data in the cost function, and the impact from missing data can be as small as a second order effect.

Regression

Regression imputation has the opposite problem of mean imputation. A regression model is estimated to predict observed values of a variable based on other variables, and that model is then used to impute values in cases where the value of that variable is missing. In other words, available information for complete and incomplete cases is used to predict the value of a specific variable. Fitted values from the regression model are then used to impute the missing values. The problem is that the imputed data do not have an error term included in their estimation, thus the estimates fit perfectly along the regression line without any residual variance. This causes relationships to be over identified and suggest greater precision in the imputed values than is warranted. The regression model predicts the most likely value of missing data but does not supply uncertainty about that value.

Stochastic regression was a fairly successful attempt to correct the lack of an error term in regression imputation by adding the average regression variance to the regression imputations to introduce error. Stochastic regression shows much less bias than the above-mentioned techniques, but it still missed one thing – if data are imputed then intuitively one would think that more noise should be introduced to the problem than simple residual variance.[5]

Multiple imputation

In order to deal with the problem of increased noise due to imputation, Rubin (1987)[10] developed a method for averaging the outcomes across multiple imputed data sets to account for this. All multiple imputation methods follow three steps.[3]

1. Imputation – Similar to single imputation, missing values are imputed. However, the imputed values are drawn *m* times from a distribution rather than just once. At the end of this step, there should be *m* completed datasets.
2. Analysis – Each of the *m* datasets is analyzed. At the end of this step there should be *m* analyses.
3. Pooling – The *m* results are consolidated into one result by calculating the mean, variance, and confidence interval of the variable of concern[11][12] or by combining simulations from each separate model.[13]

Just as there are multiple methods of single imputation, there are multiple methods of multiple imputation as well. One advantage that multiple imputation has over the single imputation and complete case methods is that multiple imputation is flexible and can be used in a wide variety of scenarios. Multiple imputation can

be used in cases where the data are missing completely at random, missing at random, and even when the data are missing not at random. A popular approach is multiple imputation by chained equations (MICE), also known as "fully conditional specification" and "sequential regression multiple imputation."[14] MICE is designed for missing at random data, though there is simulation evidence to suggest that with a sufficient number of auxiliary variables it can also work on data that are missing not at random. However, MICE can suffer from performance problems when the number of observation is large and the data have complex features, such as nonlinearities and high dimensionality.

More recent approaches to multiple imputation use machine learning techniques to improve its performance. MIDAS (Multiple Imputation with Denoising Autoencoders), for instance, uses denoising autoencoders, a type of unsupervised neural network, to learn fine-grained latent representations of the observed data.[15] MIDAS has been shown to provide accuracy and efficiency advantages over traditional multiple imputation strategies.

As alluded in the previous section, single imputation does not take into account the uncertainty in the imputations. After imputation, the data is treated as if they were the actual real values in single imputation. The negligence of uncertainty in the imputation can lead to overly precise results and errors in any conclusions drawn.[16] By imputing multiple times, multiple imputation accounts for the uncertainty and range of values that the true value could have taken. As expected, the combination of both uncertainty estimation and deep learning for imputation is among the best strategies and has been used to model heterogeneous drug discovery data.[17][18]

Additionally, while single imputation and complete case are easier to implement, multiple imputation is not very difficult to implement. There are a wide range of statistical packages in different statistical software that readily performs multiple imputation. For example, the MICE package allows users in R to perform multiple imputation using the MICE method.[19] MIDAS can be implemented in R with the rMIDAS package and in Python with the MIDASpy package.[15]

See also

- Bootstrapping (statistics)
- Censoring (statistics)
- Expectation–maximization algorithm
- Geo-imputation
- Interpolation
- Matrix completion

References

1. Barnard, J.; Meng, X. L. (1999-03-01). "Applications of multiple imputation in medical studies: from AIDS to NHANES". *Statistical Methods in Medical Research*. **8** (1): 17–36. doi:10.1177/096228029900800103 (https://doi.org/10.1177%2F096228029900800103). ISSN 0962-2802 (https://www.worldcat.org/issn/0962-2802). PMID 10347858 (https://pubmed.ncbi.nlm.nih.gov/10347858). S2CID 11453137 (https://api.semanticscholar.org/CorpusID:11453137).
2. Gelman, Andrew, and Jennifer Hill. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, 2006. Ch.25

3. Lall, Ranjit (2016). "How Multiple Imputation Makes a Difference" (https://www.cambridge.org/core/journals/political-analysis/article/how-multiple-imputation-makes-a-difference/8C6616B679EF8F3EB0041B1BC88EEBB9). *Political Analysis*. **24** (4): 414–433. doi:10.1093/pan/mpw020 (https://doi.org/10.1093%2Fpan%2Fmpw020).
4. Kenward, Michael G (2013-02-26). "The handling of missing data in clinical trials" (https://semanticscholar.org/paper/964403060982c44cc10842084105de256876b8c6). *Clinical Investigation*. **3** (3): 241–250. doi:10.4155/cli.13.7 (https://doi.org/10.4155%2Fcli.13.7). ISSN 2041-6792 (https://www.worldcat.org/issn/2041-6792).
5. Enders, C. K. (2010). *Applied Missing Data Analysis*. New York: Guilford Press. ISBN 978-1-60623-639-0.
6. Molnar, Frank J.; Hutton, Brian; Fergusson, Dean (2008-10-07). "Does analysis using "last observation carried forward" introduce bias in dementia research?" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2553855). *Canadian Medical Association Journal*. **179** (8): 751–753. doi:10.1503/cmaj.080820 (https://doi.org/10.1503%2Fcmaj.080820). ISSN 0820-3946 (https://www.worldcat.org/issn/0820-3946). PMC 2553855 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2553855). PMID 18838445 (https://pubmed.ncbi.nlm.nih.gov/18838445).
7. Kalton, Graham (1986). "The treatment of missing survey data". *Survey Methodology*. **12**: 1–16.
8. Kalton, Graham; Kasprzyk, Daniel (1982). "Imputing for missing survey responses" (https://web.archive.org/web/20200212025249/https://pdfs.semanticscholar.org/58f9/8fcc52333348a63b9e6dd5fabbdcc6fefe0e.pdf) (PDF). *Proceedings of the Section on Survey Research Methods*. American Statistical Association. **22**. S2CID 195855359 (https://api.semanticscholar.org/CorpusID:195855359). Archived from the original (https://pdfs.semanticscholar.org/58f9/8fcc52333348a63b9e6dd5fabbdcc6fefe0e.pdf) (PDF) on 2020-02-12.
9. Ren, Bin; Pueyo, Laurent; Chen, Christine; Choquet, Elodie; Debes, John H; Duchene, Gaspard; Menard, Francois; Perrin, Marshall D. (2020). "Using Data Imputation for Signal Separation in High Contrast Imaging". *The Astrophysical Journal*. **892** (2): 74. arXiv:2001.00563 (https://arxiv.org/abs/2001.00563). Bibcode:2020ApJ...892...74R (https://ui.adsabs.harvard.edu/abs/2020ApJ...892...74R). doi:10.3847/1538-4357/ab7024 (https://doi.org/10.3847%2F1538-4357%2Fab7024). S2CID 209531731 (https://api.semanticscholar.org/CorpusID:209531731).
10. Rubin, Donald (9 June 1987). *Multiple imputation for nonresponse in surveys*. Wiley Series in Probability and Statistics. Wiley. doi:10.1002/9780470316696 (https://doi.org/10.1002%2F9780470316696). ISBN 9780471087052.
11. Yuan, Yang C. (2010). "Multiple imputation for missing data: Concepts and new development" (https://support.sas.com/rnd/app/stat/papers/multipleimputation.pdf) (PDF). *SAS Institute Inc., Rockville, MD*. **49**: 1–11.
12. Van Buuren, Stef (2012-03-29). "2. Multiple Imputation". *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics Series. Vol. 20125245. Chapman and Hall/CRC. doi:10.1201/b11826 (https://doi.org/10.1201%2Fb11826). ISBN 9781439868249.
13. King, Gary; Honaker, James; Joseph, Anne; Scheve, Kenneth (March 2001). "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation" (https://www.cambridge.org/core/journals/american-political-science-review/article/analyzing-incomplete-political-science-data-an-alternative-algorithm-for-multiple-imputation/9E712982CCE2DE79A574FE98488F212B). *American Political Science Review*. **95** (1): 49–69. doi:10.1017/S0003055401000235 (https://doi.org/10.1017%2FS0003055401000235). ISSN 1537-5943 (https://www.worldcat.org/issn/1537-5943).

14. Azur, Melissa J.; Stuart, Elizabeth A.; Frangakis, Constantine; Leaf, Philip J. (2011-03-01). "Multiple imputation by chained equations: what is it and how does it work?" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241). *International Journal of Methods in Psychiatric Research*. **20** (1): 40–49. doi:10.1002/mpr.329 (https://doi.org/10.1002%2Fmpr.329). ISSN 1557-0657 (https://www.worldcat.org/issn/1557-0657). PMC 3074241 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241). PMID 21499542 (https://pubmed.ncbi.nlm.nih.gov/21499542).
15. Lall, Ranjit; Robinson, Thomas (2021). "The MIDAS Touch: Accurate and Scalable Missing-Data Imputation with Deep Learning" (https://www.cambridge.org/core/journals/political-analysis/article/abs/midas-touch-accurate-and-scalable-missingdata-imputation-with-deep-learning/5007854F57E88AF16D69BCCA4C5AF1FF). *Political Analysis*. doi:10.1017/pan.2020.49 (https://doi.org/10.1017%2Fpan.2020.49).
16. Graham, John W. (2009-01-01). "Missing data analysis: making it work in the real world". *Annual Review of Psychology*. **60**: 549–576. doi:10.1146/annurev.psych.58.110405.085530 (https://doi.org/10.1146%2Fannurev.psych.58.110405.085530). ISSN 0066-4308 (https://www.worldcat.org/issn/0066-4308). PMID 18652544 (https://pubmed.ncbi.nlm.nih.gov/18652544).
17. Irwin, Benedict (2020-06-01). "Practical Applications of Deep Learning to Impute Heterogeneous Drug Discovery Data". *Journal of Chemical Information and Modeling*. **60** (6): 2848–2857. doi:10.1021/acs.jcim.0c00443 (https://doi.org/10.1021%2Facs.jcim.0c00443). PMID 32478517 (https://pubmed.ncbi.nlm.nih.gov/32478517).
18. Whitehead, Thomas (2019-02-12). "Imputation of Assay Bioactivity Data Using Deep Learning". *Journal of Chemical Information and Modeling*. **59** (3): 1197–1204. doi:10.1021/acs.jcim.8b00768 (https://doi.org/10.1021%2Facs.jcim.8b00768). PMID 30753070 (https://pubmed.ncbi.nlm.nih.gov/30753070).
19. Horton, Nicholas J.; Kleinman, Ken P. (2007-02-01). "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839993). *The American Statistician*. **61** (1): 79–90. doi:10.1198/000313007X172556 (https://doi.org/10.1198%2F000313007X172556). ISSN 0003-1305 (https://www.worldcat.org/issn/0003-1305). PMC 1839993 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839993). PMID 17401454 (https://pubmed.ncbi.nlm.nih.gov/17401454).

External links

- Missing Data: Instrument-Level Heffalumps and Item-Level Woozles (https://archive.today/20130223193833/http://division.aomonline.org/rm/1999_RMD_Forum_Missing_Data.htm)
- Multiple-imputation.com (https://web.archive.org/web/20120831160303/http://www.multiple-imputation.com/)
- Multiple imputation FAQs, Penn State U (https://web.archive.org/web/20050212022244/http://www.stat.psu.edu/~jls/mifaq.html)
- A description (http://www.stat.fi/isi99/proceedings/arkisto/varasto/scho0502.pdf) of hot deck imputation from Statistics Finland.
- Paper (https://web.archive.org/web/20160303174300/http://www.amstat.org/sections/srms/Proceedings/papers/1993_005.pdf) extending Rao-Shao approach and discussing problems with multiple imputation.
- Paper (http://www.iaeng.org/publication/WCE2012/WCE2012_pp391-394.pdf) Fuzzy Unordered Rules Induction Algorithm Used as Missing Value Imputation Methods for K-Mean Clustering on Real Cardiovascular Data.
- [1] (http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/data-editing-and-imputation/index.html) Real world application of Imputation by the UK Office of